Title: Changes in General and Specific Psychopathology Factors Over A Psychosocial Intervention

Running title: Psychopathology Factors in Psychotherapy

Author names and affiliations:

Matthew P. Constantinou, MPhil, Ian M. Goodyer, MD FMedSci, Ivan Eisler, PhD FAcSS, Stephen Butler, PhD, Abdullah Kraam, MD FRCPsych, Stephen Scott, FRCPsych FMedSci, Stephen Pilling, PhD, Elizabeth Simes, MA, Rachel Ellison, BSc, Elizabeth Allison, DPhil, and Peter Fonagy, PhD FMedSci

Matthew Constantinou (m.constantinou@ucl.ac.uk), Elizabeth Simes (e.simes@ucl.ac.uk), Rachel Ellison (rachelmaryellison@gmail.com), Dr Elizabeth Allison (e.allison@ucl.ac.uk), Professor Stephen Pilling (s.pilling@ucl.ac.uk), and Professor Peter Fonagy (p.fonagy@ucl.ac.uk) are from the Research Department of Clinical, Educational and Health Psychology, Division of Psychology and Language Sciences, University College London, London, UK.

Professor Ivan Eisler (ivan.eisler@kcl.ac.uk) is from the South London and Maudsley NHS Foundation Trust and the Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.

Professor Stephen Scott (stephen.scott@kcl.ac.uk) is from the Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.

Professor Ian Goodyer (ig104@cam.ac.uk) is from the Department of Psychiatry, University of Cambridge, Cambridge, UK.

Professor Stephen Butler (stbutler@upei.ca) is from the Department of Psychology, University of Prince Edward Island, Charlottetown, Canada.

Dr Abdullah Kraam (a.kraam@leeds.ac.uk) is from the University of Leeds and Leeds Community Healthcare NHS Trust, Leeds, UK.

Corresponding author: Correspondence should be sent to Matthew Constantinou, University College London, 1-19 Torrington Place, London, WC1E 7HB. E-mail: m.constantinou@ucl.ac.uk. Tel: +44 (0)20 7679 3406.

Manuscript Funding

This study, which is a secondary analysis of a registered clinical trial (ISRCTN77132214), was supported by a studentship from the UK Medical Research Council (MR/J500422/1) awarded to Matthew Constantinou. Funders had no role in the secondary analysis or interpretation of findings.

The primary study received funding from the Department for Children, Schools and Families and the Department of Health and Social Care. Non-financial support was provided by the Youth Justice Board and National Institute for Health Research (NIHR) Clinical Research Network during the conduct of the primary study.

Acknowledgements

The authors would like to thank Nicole Hickey, MSc, of Imperial College London, who prepared the youth offending data from official records from the Police National Computer and Youth Offender Information System for all nine MST sites. The authors also

acknowledge the excellent work of the team of research assistants at University College London, University of Leeds, and University of Cambridge involved in acquiring the data across the nine MST sites.

Financial Disclosure Statements

Professor Peter Fonagy is in receipt of a National Institute for Health Research (NIHR) Senior Investigator Award (NF-SI-0514-10157), and was in part supported by the NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames at Barts Health NHS Trust.

Professor Ian M. Goodyer is supported by a Wellcome Trust Strategic Award (reference: 095844) and receives consulting fees from Lundbeck.

Professor Ivan Eisler was in part supported by NIHR grants from Health Technology Assessment and Health Service Delivery Research programmes

Dr Abdullah Kraam is anticipating to be in receipt of RfPB funding as Co-CI. No other funding or conflicts of interest to declare.

Professor Stephen Scott is principal or co-investigator on 4 grants from the UK National Institute of Health Research and 1 from the UK Medical Research Council to his employer, Kings College London.

Professor Stephen Pilling is supported by the National Institute for Health Research University College London Hospital′s Biomedical Research Centre and the National Institute of Clinical Excellence.

Professor Stephen Butler, Dr Elizabeth Allison, Matthew Constantinou, Rachel Ellison, and Elizabeth Simes report no outstanding financial interests or potential conflicts of interest.

Keywords: Bifactor, p factor, general psychopathology, psychotherapy, intervention.

Social Media

Facebook: New Study @JAACAP using @UCL START trial data finds changes in both general and specific psychopathology factors over a psychosocial intervention for adolescent antisocial behavior #UCL_PSA #mentalhealth #UCLSTARTtrial https://www.ucl.ac.uk/psychoanalysis/research/systemic-therapy-risk-teens-start

Twitter: New Study @JAACAP using @UCL START trial data finds changes in both general and specific psychopathology factors over a psychosocial intervention for adolescent antisocial behavior @UCL_PSA @mpconstantinou @PeterFonagy #mentalhealth #UCLSTARTtrial

Abstract

Objective: Recent research suggests that comorbidity among child and adolescent psychiatric symptoms can be explained by a single general psychopathology ('p') factor, as well as more specific factors summarizing clusters of symptoms. We investigated within- and between-person changes in the general and specific psychopathology factors over a psychosocial intervention.

Method: We ran a secondary analysis of the Systemic Therapy for At-Risk Teens study, a pragmatic randomized controlled trial that compared the effects of multisystemic therapy to management-as-usual for reducing antisocial behavior in 684 adolescents (82% male; 11-18 at baseline) over an 18-month period. The general p factor, as well as specific antisocial, attention, anxiety, and mood factors, were estimated from a symptom-level analysis of a set of narrow-band symptom scales measured repeatedly over the study. General and specific psychopathology factors were assessed for reliability, validity, and within- and between-person change using a parallel process multilevel growth model.

Results: A revised bifactor model that included a general p factor and specific anxiety, mood, antisocial, and attention factors with cross-loadings fit the data best. While the factor structure was multidimensional, p accounted for most of the variance in total scores. The p, anxiety, and antisocial factors predicted within-person variation in external outcomes. p and antisocial factors showed within-person reductions, while anxiety showed within-person increases over time. Despite individual variation in baseline factor scores, adolescents showed similar rates of change.

Conclusion: The bifactor model is useful for teasing apart general and specific therapeutic changes which are conflated in standard analyses of symptom scores.

Clinical trial registration information: START (Systemic Therapy for At Risk Teens): A National Randomised Controlled Trial to Evaluate Multisystemic Therapy in the UK Context. http://www.isrctn.com; ISRCTN77132214

Keywords: Bifactor, p factor, general psychopathology, psychotherapy, intervention

Introduction

Clinical researchers typically assess therapeutic change through child, carer, or teacher reports of disorder-specific symptoms. For instance, a psychological intervention for social phobia is deemed successful if social phobia symptoms decline from pre- to post-treatment, typically below a clinical threshold. However, this assumes that psychiatric disorders are independent of each other and can each be reliably measured. Yet symptoms from various disorders, as well as disorders themselves, co-occur more strongly than expected by chance,[1,2] which points to broader underlying processes.[3]

Recognising the comorbidity among symptoms, clinical researchers also assess therapeutic change through broadband measures, such as internalizing (a composite of depressive, somatic, and anxiety symptoms) and externalizing (a composite of attentional, behavioural, and substance-use problems).[4] Recently, a single 'general psychopathology' factor, or p factor, has been shown to summarize the co-occurrences among many symptoms of differing types, in addition to more specific factors like internalizing and externalizing.[5,6] Put differently, all symptoms are thought to share something in common, as well as things unique to subsets of disorders.[7]

A growing number of studies support the bifactor model–which includes both general and specific psychopathology factors–as a candidate structure of psychopathology in children and adolescents.[8-15] However, the benefits of analyzing treatment outcomes with the bi-factor model have yet to be shown. For example, the p factor may capture common therapeutic effects,[7] which by definition affect all symptoms non-specifically.[16] Because specific factors are residualized for the common variance, they may be useful in identifying the specific effects of a treatment on the problems it was designed to engage with. This is difficult to achieve with total and subscale scores because their variances are not independent. In turn,

researchers may identify changes in disorder-specific subscales that are actually the result of common effects.[17] Moreover, most work on the bifactor model is based on between-person differences. It is unknown whether symptoms positively co-occur at the within-person level. Such evidence would provide new insight into the way in which therapeutic change occurs.

In the current study, we analyzed data from the Systemic Therapy for At-Risk Teens (START) trial which compared the effects of multisystemic therapy to management-as-usual in reducing antisocial behaviour. Treatment arms did not differ but there were widespread changes in self-reported emotional and behavioural problems.[18] We first determined whether the bifactor model adequately captured within-person associations among self-reported emotional and behavioral symptoms over the study period, collapsed across treatment arms. We then assessed the general and specific factors for their reliability using model-based reliability estimates, and concurrent validity using external outcomes of criminal activity and academic attendance. Finally, we assessed within-person change in the general and specific psychopathology factors, as well as between-person differences in within-person change, using a multilevel growth model.

Method

Trial Design

Full details of the START trial can be found in the study publication.[18] START was a pragmatic individually randomised multicentre superiority trial which compared the effects of multisystemic therapy followed by management-as-usual to management-as-usual alone in reducing out-of-home placements and criminal activity in adolescents with moderate to severe conduct problems. Assessment occurred at baseline, post-treatment (6 months), follow-up 1 (12-months), and follow-up 2 (18 months).

Participants

Eligible adolescents met at least one of the following criteria: 1) persistent (weekly) and enduring (≥ 6 months) violent and aggressive interpersonal behavior; 2) at least one conviction plus three additional warnings, reprimands, or convictions; 3) a current DSM-IV diagnosis of CD that had not responded to treatment; 4) a permanent school exclusion for antisocial behavior; 5) a significant risk of harm to others or self. Eligible adolescents also met at least three severity criteria indicative of past difficulties across several settings (e.g. school non-attendance or exclusion, offending, child protection investigation, high risk of coming into care). As this was a pragmatic trial, participants were not excluded for comorbid disorders except for psychosis, acute suicidality, and generalized learning difficulties.

Adolescents were referred from social services, youth justice, schools, child and adolescent mental health services, and voluntary services. The final sample consisted of 684 adolescents (M_{age} = 13.8, SD_{age} = 1.4, 11-18 at baseline), the majority of whom were male (82%), Caucasian (78%), and of a low-moderate socio-economic background (77% on state benefits). Most adolescents had a diagnosis of conduct disorder (78%) or any conduct disorder (81%). Other frequent disorders included attention-deficit hyperactivity disorder-combined (30%) and any emotional disorder (24%). The sample's diagnostic profile and further details of referral pathways can be found in trial paper.[18] Written consent was obtained from all participants, and the study protocol was approved by the London South-East Research Ethics Committee (09/H1102/55).

Intervention and Randomization

Multisystemic therapy (MST) is a family-based intervention which targets the multiple systems influencing chronic and pervasive antisocial behavior in adolescents, including the home, school, and peer environments.[19] This is primarily achieved through caregivers who are taught how to enhance family relationships via communication skills and

parenting techniques, and how to encourage school attendance and achievement rather than delinquent peer activity. Techniques from cognitive-behavioral therapy, behavioral parent training, and pragmatic family therapy are integrated and tailored to the needs of each family. There were nine MST pilot sites with at least 12 months experience of running the programme.

Management-as-usual (MAU) replicated best-practice in managing the complex needs of antisocial youth in community settings. Interventions based on treatment guidelines were administered on an ad-hoc basis (e.g., support to re-engage with education, anger management, victim awareness programmes). MAU was multi-component and no less intensive than MST; the main differences were that MAU lacked standardization, an overarching formulation of the problem, and weekly expert supervision.

Adolescents were randomised to MST or MAU by an equal allocation ratio using stochastic minimisation, balancing for treatment centre, sex, current age (<15 or ≥15), and age at onset of antisocial behavior (≤11 or >11).

Measures

All measures were taken at baseline, post-treatment, follow-up 1, and follow-up 2. Emotional and behavioral problems were assessed using the Strengths and Difficulties Questionnaire (SDQ).[20-22] We used child-reported items from the emotional problems, conduct problems, and attention-hyperactivity subscales in our measurement model, each of which has five items rated on a three-point scale (not true, somewhat true, certainly true). Items from the peer problems and prosocial subscales were not included because we aimed to limit our analysis to psychiatric symptoms, which naturally excludes prosocial items, and general difficulties engaging with peers are not symptoms in and of themselves. That is, interpersonal problems reflect a broader level of analysis (e.g., a child may be bullied because

they appear nervous and withdrawn or because they are bold and irritable), which have not yet been thoroughly validated in the bifactor model of psychopathology.

We also used the Mood and Feelings Questionnaire-Short Form (MFQ) to increase our internalizing item pool. The MFQ is a 13-item measure shown to reliably assess depression in young people,[23-25] with items scored on a three-point scale (not true, somewhat true, true). Past research suggests that the MFQ captures a single between-person depression factor,[26,27] but our exploratory within-level factor analysis revealed two clear factors (see Table S1, available online). The first factor reflected problems with self-attitudes and the second captured problems in mood. We included the top-five loading items of the mood factor to balance the internalizing and externalizing content of the SDQ, and ensure that the p factor was not biased to any given symptom domain. We used five items to ensure that equal numbers of items loaded on each factor. The mood factor was used because the self-attitudes factor reflects a transdiagnostic construct that has not yet been validated in the bifactor model of psychopathology.

We obtained official records of violent and non-violent offences committed over the study period from the Police National Computer and Young Offender Information System, and records of the number of school exclusions from the National Pupil Database.

Statistical Analysis

Model Comparison

Using the SDQ and MFQ item-level data, we estimated three multi-level confirmatory factor analysis (CFA) models–a correlated traits model, common factor model, and bifactor model–and compared their ability to summarize within-person variation in emotional and behavioral problems over the study. We arranged the repeated observations for each item in long format and listed each item in wide format for computational ease (see Supplement 1 for

further details, available online). We also collapsed across treatment arms to ensure adequate power. Longitudinal multi-level CFA and single-level CFA differ in how factors are estimated. In a single-level CFA with wide-formatted data, a factor is repeatedly estimated at each time-point. Furthermore, a factor loading reflects the way in which an item is predicted to co-occur with other items between individuals at a given time-point. In contrast, multi-level CFAs estimate a single factor across time-points at the within-level, and differences between subjects at the between-level. A factor loading reflects the way in which an item is predicted to co-occur with other items over time for each individual.

Only factors at the within-level were estimated as we were interested in summarizing the covariation among symptoms for each individual over time, rather than teasing apart the within- and between-level factor structures (see Supplement 1 and Figure S1, available online). Furthermore, within- and between-person variances were subsequently estimated in a multi-level growth model to investigate between-person differences in within-person change. Within-level correlations between the general and specific factors were constrained to zero, as well as the correlations between specific factors. Item overlap was accounted for by correlating the residuals for SDQ item 13 ('I am often unhappy') with MFQ item 1 ('I felt miserable/unhappy'), and SDQ item 2 ('I am restless') with MFQ item 4 ('I was very restless').

An advantage of multi-level CFA is that fewer parameters are needed to estimate growth in complex models since the analysis is collapsed over time.[28] A disadvantage is that it is not possible to test for conventional measurement invariance, i.e. the extent to which within-person change is driven by changes in measurement properties (e.g. differential item functioning or response biases) rather than the factors. Instead, parameters are assumed to be invariant and are modelled as such. We still tested for measurement invariance using the

conventional method, but the results are not directly translatable to the multi-level approach (see Supplement 2, available online).

Models were estimated using the weighted least squares means and variances adjusted estimator (WLSMV) since it is designed for non-normal and categorical indictors.[29] Overall fit was assessed using the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Acceptable and excellent fit, respectively, were defined by CFI values ≥ .90 and ≥ .95, TLI values ≥ .90 and ≥ .95, RMSEA values ≤ .08 and ≤ .06, and SRMR values ≤ .08 and ≤ .06.[30] Models could not be compared using chi-square values because they were not nested. Therefore, we (cautiously) adopted the guidelines for comparing nested models: increases in CFI > ~.01 (and by generalization, TLI > ~.01), and decreases in RMSEA > ~.015 (and by generalization, SRMR > ~.015) between the more and less restricted models indicated a meaningful improvement in fit.[31] We also re-estimated models using the robust maximum-likelihood estimator to compare their fit using the Akaike Information Criteria (AIC) and sample-size adjusted Bayesian Information Criteria (BIC). Bifactor models have a tendency to overfit noise, so it is important to include fit statistics which penalize for model complexity.[32] A difference of 2 (AIC/BIC) between models was considered negligible; a difference of 2-7 (AIC) or 2-6 (BIC) suggested some evidence favouring the competing model; a difference of 7-10 (AIC) or 6-10 (BIC) suggested strong evidence favouring the competing model, and a difference greater than 10 (AIC/BIC) suggested very strong evidence favouring the competing model.[33]

Factor Score Estimation

We estimated within-level Bayesian plausible values (BPVs) for each individual at each time-point for the p factor and specific anxiety, mood, antisocial, and attention factors.

Like factor scores, BPVs are observed estimates of latent variables, but unlike factor scores, they also take into account the uncertainty or 'indeterminacy' inherent in estimating factor scores by averaging over a distribution of possible factor scores using multiple imputation.[34,35] Theoretical and simulation studies suggest that BPVs provide less biased estimates of population parameters than factor scores.[34-37] In practice, BPVs and factor scores likely produce similar estimates when sample sizes are sufficient.[38,39] We used BPVs because bifactor models that are 'essentially unidimensional' (i.e. show a strong general factor and small but salient specific factors[40]) produce specific factors with relatively low reliability. It is thus important to statistically incorporate estimates of imprecision when using factor scores in secondary analyses.[41,42]

BPVs were estimated using the same multi-level growth model in the main analysis (see below) to minimise bias (see Table S2 for correlations between BPVs).[34] There is little consensus over how many imputations to estimate. While Asparouhov and Muthén suggest that five imputations is sufficient for secondary analyses (unpublished manuscript, August 21, 2010) we estimated one-hundred imputations with a thinning rate of 1 (e.g., random estimates were sampled on every iteration).

Factor Score Evaluation

Reliability

In addition to incorporating reliability estimates in the growth curve analysis using BPVs, we calculated model-based reliability estimates using omega hierarchical (ω_H), omega hierarchical subscale (ω_{Hs}), explained common variance (ECV), and explained common variance subscale (ECV_s). The ω_H and ECV are estimates of the amount of test variance and common variance explained by a general factor, respectively (or by a specific factor if using ω_{Hs} and ECV_s).[43] We also determined the total information functions for factor scores based

on WLSMV estimates of the chosen bifactor model. “Information” is the inverse of measurement error and reflects the amount of precision or reliability in factor scores.[44] Information varies at different estimates of the latent trait; more information indicates higher measurement precision for a given factor score.

Validity

We assessed the validity of within-person variation in the general and specific factor BPVs by regressing two external outcomes–official records of the number of offences and school exclusions–onto BPVs at the within-level. We also assessed the validity of within-person change in general and specific BPVs by regressing each external outcome onto each factor, a linear time variable (which captures change in the external outcome over time), and time*factor interactions (which captures how changes in the factors over time predict changes in the external outcomes) using Poisson multilevel growth models.

Multilevel Growth Model

We analyzed within-person changes in BPVs, as well as between-person differences in within-person change, using a parallel process multilevel growth model. We arranged the repeated observations for each factor in long format and listed each factor in wide format (see Supplement 1 and Figure S2, available online). Within-level BPVs for each factor were regressed in parallel onto linear and quadratic time variables. Random intercepts, random linear slopes, and random quadratic slopes were estimated for each factor at the between-level, and covaried for each factor and across factors. Random effects were controlled for individual differences in baseline age (centered) (see Supplement 1 for model details, available online).

We evaluated residuals at the within- and between-levels for deviations from normality and homoscedasticity, and estimated the model using the robust maximum

likelihood estimator. Partially standardized regression coefficients were analyzed with two-tailed Wald tests (BPVs are standardized, e.g., have a mean of 0 and variance of 1, hence regression coefficients are partially standardized on the Y-axis). Mean changes in BPVs should thus be interpreted as the change in standardized units of the factors with an increase in time, assuming equal item thresholds over time when factors are held constant and equal loadings.

Dropout was the most common cause of missing values, with 12% of cases dropping out at 6 months, 12% at 12 months, and 19% at 18 months. Pattern-mixture models supported the assumption that the unobserved data were missing at random. Missing data were thus handled with full-information maximum likelihood. All analyses were run in Mplus 8.0.[45]

Results

Model Comparison

Table S3 (available online) presents an item-level polychoric correlation matrix. The common factor model included a single within-level factor upon which all items loaded, and fit the data poorly (CFI = .69, TLI = .65, RMSEA = .12, SRMR = .11; see Table S4 for factor loadings, available online). The correlated factors model included four within-level factors which each reflected an SDQ or MFQ subscale: anxiety, mood, antisocial, and attention. The correlated factors model approached an acceptable fit (e.g., CFI = .89, TLI = .87; RMSEA = .07, SRMR = .07). Furthermore, the correlated factors model fit better than the common factor model (ΔCFI = .24, ΔTLI = .18, ΔRMSEA = .05, ΔSRMR = .04, ΔAIC = 2047, ΔBIC = 2032). Table S5 (available online) provides factor loadings and correlations for the correlated factors model.

The bifactor model included a general (‘p’) factor upon which all items loaded, as well as four specific factors (anxiety, mood, antisocial, and attention). Like the correlated

factors model, fit indices were almost acceptable (CFI = .86, TLI = .82, RMSEA = .08, SRMR = .07; see Table S6 for factor loadings, available online). Since Mplus does not provide modification indices for multi-level factor analysis, we examined the factor loadings of a multi-level bifactor EFA for substantial ($\lambda \geq .32$)[46] and theoretically plausible cross-loadings (see Table S7, available online). Three items, two from the attention factor and one from the antisocial factor, cross-loaded onto the anxiety factor: SDQ item 7 ('I [do not] usually do as I am told', $\lambda = -.46$), 21 ('I [do not] think before I do things', $\lambda = -.33$) and 25 ('I [do not] finish the work I am doing', $\lambda = -.33$). These items reflect behavioral control which positively co-occurs with internalizing problems after controlling for general psychopathology.[47,48] Moreover, SDQ item 16 from the anxiety factor ('I am nervous in new situations') negatively cross-loaded onto the antisocial factor ($\lambda = -.32$), which is expected if the specific antisocial factor overlaps with fearlessness.[49]

A revised bifactor model which included these cross-loadings fit the data well (CFI = .93, TLI = .91, RMSEA = .06, SRMR = .05), and better than the standard bifactor model (ΔCFI = .07, ΔTLI = .09, ΔRMSEA = .02, ΔSRMR = .02, ΔAIC = 429**,** ΔBIC = 419) and correlated factors model (ΔCFI = .04, ΔTLI = .04, ΔRMSEA = .01, ΔSRMR = .02, ΔAIC = 415, ΔBIC = 370). Therefore, BPVs were estimated for the revised bifactor model (see Table 1 for standardized loadings).

Factor Score Evaluation

Reliability

Figure 1 shows the total information functions for the general and specific psychopathology factors. The p factor had a larger information function than the specific factors, since it had more items with larger average loadings. Factor scores were thus most reliable for the p factor, particularly at the average level. Omega hierarchical estimates were

similar: the p factor showed a relatively large omega hierarchical (ω_H = .71). Hence, 71% of the overall variance, and 78% of the reliable variance (ω_H/ω) in total raw scores was accounted for by the p factor. In contrast, ω_{Hs} for the specific factors ranged from .32-.43 (see Table 1). Nonetheless, the common variance was equally split between the general (ECV = .50) and specific (ECV_s = .50) factors, demonstrating that a factor structure can still be multidimensional, even if total scores primarily reflect a single dimension, or factor scores most reliably measure the general factor.[40]

Validity

Within-person variability in antisocial BPVs positively predicted variability in the number of offences committed (β = .12, *p* = .043, 95% CI [.01, .24]). That is, higher (or lower) antisocial scores co-occurred with more (or less) offences for each individual over time. Moreover, within-person variability in anxiety BPVs negatively predicted variability in the number of exclusions (β = -.13, *p* = .040, 95% CI [-.25, -.01]). Therefore, higher (or lower) anxiety scores co-occurred with less (or more) exclusions for each individual over time.

The number of school exclusions significantly declined over time (β = -.14, *p* = .037, 95% CI [-.26, -.01]), as well as the number of offences, albeit marginally (β = -.11, *p* = .052, 95% CI [-.21, .00]). The only factor whose effect over time predicted within-person changes in the external outcomes was p. Specifically, reductions in p (see 'Growth Model' for slope) positively predicted the number of offences committed (β = .13, *p* = .012, 95% CI [.03, .22]). In other words, decreases in p predicted decreases in offences. Moreover, reductions in p marginally predicted the number of exclusions (β = .06, *p* = .064, 95% CI [.00, .22]). That is, decreases in p marginally predicted decreases in school exclusions.

Multilevel Growth Model

Figure 2 shows the observed and predicted within-level growth curves pooled across individuals for the p factor and specific anxiety, mood, antisocial, and attention factor BPVs. Both the p factor (β = -.28, $p < .001$, 95% CI [-.41, -.16]) and specific antisocial factor (β = -.27, $p = .002$, 95% CI [-.43, -.10]) decreased over time for each individual (see Figure 2a). Furthermore, the p factor (β = .03, $p = .087$, 95% CI [-.01, .07]), but not the antisocial factor (β = .01, $p = .836$, 95% CI [-.04, .05]), showed a marginally significant quadratic growth term. That is, within-person decline in the p factor decelerated towards the follow-up period. The specific anxiety factor showed a significant linear increase over time for each individual (β = .17, $p = .021$, 95% CI [.03, .32]), which occurred at a steady pace (quadratic trend: β = .00, $p = .984$, 95% CI [-.05, .05]; see Figure 2b). Finally, the specific mood factor did not deviate from baseline (β = -.06, $p = .42$, 95% CI [-.20, .08]), while the specific attention factor maintained an elevated level throughout (linear slope: β = -.01, $p = .89$, 95% CI [-.12, .14], See Figure 2c). Results were similar when cross-loadings were removed from the initial factor model, but within-person decline in the antisocial factor was no longer significant (see Supplement 3 and Figure S3a, available online).

Adolescents significantly varied in their initial factor levels, but not in how they changed over time (see Table S8, available online). Moreover, the correlations among random effects within and between factors, and the effects of baseline age on the random effects, were not significant, mainly due to the increase in standard errors using BPVs (see Table S8 and S9, available online).

Discussion

We aimed to tease apart the general and specific aspects of emotional and behavioral change in adolescents over a psychosocial intervention for antisocial activity. We found that within-person associations between symptoms were best modeled with a multilevel bifactor

model that included a general p factor and specific anxiety, mood, antisocial, and attention factors, as well as cross-loadings. Despite this multidimensional factor structure, p factor scores showed high reliability while specific factor scores were only modest. Nonetheless, both general p and specific antisocial and anxiety factors were externally validated against official records of criminal and academic activity. Within-person levels of the p factor and antisocial factor declined over the study, while the anxiety factor increased. The mood and attention factors did not change over time.

No study to our knowledge has examined within-person changes in the bifactor dimensions over a psychosocial intervention. However, one study assessed within-person changes in the bifactor dimensions of personality pathology over a ten-year naturalistic study, where patients received various treatments in an uncontrolled fashion.[28] While the populations, time-scales, treatments, and measures differ between studies, we both found that the general factor declined over time. It is tempting to argue that reductions in these general factors reflect the non-specific or universal effects of psychological therapies.[7] This might explain why seemingly specific interventions resulted in broad improvements in emotional and behavioral problems in the primary study[18]. Further, it might explain why the general personality disorder factor predicted widespread improvements in social, occupational, and recreational functioning. [28] It may even explain why different evidence-based therapies tend to result in similar outcomes,[16] as they could mainly target p.[5] Nonetheless, further controlled studies with large datasets and explicit tests for measurement invariance are necessary to test these hypothesis.

Within-person changes were also observed in the specific factors, which we argue reflect the unique effects of treatment after separating out changes common to all symptoms. Reductions in the specific antisocial factor likely reflect the specific aim of the interventions, which was to reduce antisocial behavior. This is supported by the finding that within-person

variability in the antisocial factor over time positively predicted official records of criminal activity, although the association was modest. It should be noted that decline in the antisocial factor was no longer significant when cross-loadings were removed from the model. Some may argue that decline in antisociality was a function of increases in anxiety, since the item that cross-loaded on the antisocial factor traditionally reflects anxiety (SDQ item 16: 'I am nervous in new situations'). Nonetheless, item 16 loaded more strongly onto, and hence better reflects, the antisocial factor than the anxiety factor.

Interpreting the within-person increases in anxiety is particularly interesting since standard analyses of anxiety-related problems showed a decline in the primary study.[18] One explanation is that once common therapeutic effects are controlled for, an increase in anxiety reflects a facilitative effect, whereby adolescents regained some level of fearfulness which is characteristically reduced in adolescents with severe conduct problems.[50] This is partially supported by the finding that within-person variability in anxiety scores negatively but modestly predicted official records of school exclusions over time. Others have also reported that the specific internalizing factor is positively associated with teacher-reported school functioning.[51] Alternatively, anxiety problems may have replaced antisociality as they were partial drivers of antisocial behavior. Antisocial activity can serve to protect some young people from the social situations they find challenging and will have to confront once delinquent socializing is no longer available to support their avoidance.[52] Further work is required to test these hypotheses.

The specific mood and attention factors showed little within-person change over time, yet we found substantial reductions in these problems using standard analyses of symptom subscales.[18] We thus argue that therapeutic change in these problems was secondary to more common processes captured by p. It is noteworthy that most outcome studies do not separate out the general and specific variance in symptom measures. Therefore, much of what is

reported as disorder-specific change using symptom subscales may be underpinned by common processes like reductions in general psychopathology.[17]

It is important to interpret our results within the confines of our modeling approach. We used a multi-level bifactor model to achieve stable estimates, but at the cost of longitudinal measurement invariance testing. Measurement invariance is still assumed within the parameters: an item with a strong within-level factor loading is inherently metric invariant, in that it consistently covaries with other items over time. However, we cannot determine the relative influence of method effects (e.g., differences in measurement properties over time) or heterotypic change in psychopathology (e.g., anxiety may have increased as a function of adolescence). Therefore, the extent to which within-person change in our factors solely reflects treatment effects should be interpreted with caution. We encourage future attempts at modeling growth in the bifactor dimensions to begin with a conventional single-level model, where factors are repeatedly estimated at each time-point and parameters can be explicitly tested for longitudinal measurement invariance. Results from a preliminary invariance analysis using the conventional method suggest that factor loadings and item thresholds were partially invariant (see Supplement 2, available online).

Another limitation is that we relied on Bayesian plausible values (BPVs) rather than latent variables for computational ease. While BPVs take into account the (un)reliability of factor scores, they are still an imperfect measure of latent variables. This is especially relevant when estimating BPVs for specific factors, which showed lower reliability than the general factor since they included fewer items with weaker loadings. Consequently, the variability estimated for specific factor BPVs was higher, and might have increased type II error rates in structural coefficients (e.g., the lack of significant correlations among random effects). Our growth curves may thus lack precision and require caution in their interpretation. On a related note, model-based reliability estimates assume continuous

outcome variables. The consequences of applying continuous-variable formulae to ordered categorical variables, as we have done (as well as many others), are unknown. Researchers should thus be cautious in interpreting reliability estimates for categorical outcomes in the same way as continuous outcomes.

A further issue is our inclusion of cross-loadings which makes our chosen model more exploratory, and may introduce parameter bias by violating a simple factor structure.[53] Work on the bifactor model of psychopathology is in its infancy, making it hard to rigorously substantiate cross-loadings. However, they coincide with emerging work on the relationship between specific factors and transdiagnostic mechanisms (e.g., anxiety without its pathological component may overlap with mechanisms associated with inhibition and compliance–mechanisms that predict better school functioning).[51] Validating the general and specific factor BPVs against external outcomes was important in showing that they were not merely exploratory artifacts, and we encourage future bifactor studies to include external criteria as a minimum. Nonetheless, the associations between factor scores and external outcomes were only modest, and validation does not mean causation: changes in p may not have caused changes in criminal and academic activity.

A strength of our multi-level modeling approach is that changes in general and specific psychopathology factors were analyzed at the within-person level. Most, but not all,[54] bifactor studies have modeled these factors at the between-person level. Therefore, while there is growing evidence that the average levels of symptoms co-occur *between* individuals[8-15], it was unknown whether symptoms positively co-occur *within* a given individual. Our findings suggest that individual levels of general psychopathology can be reduced by a psychosocial intervention.[7] Therefore, therapeutic change should be assessed more globally, within the system of emotional and behavioral problems experienced at the clinical or subclinical level by each young person, rather than just the disorder treated.[55]

It may be fruitful to examine other large trials using a bifactor approach to inform hypotheses about specific interventions that appear non-specifically related to multiple disorders. For example, a bifactor analysis separating out syndromal, spectral, and general factor changes may identify different predictors of outcome at different levels of the hierarchy, enabling improved answers to the 'what works for whom' question,[56] such as identifying those who are more likely to benefit from pharmacological as opposed to psychosocial interventions or indeed from different forms of psychosocial treatment. Similar considerations apply to comparisons between pharmacological agents where evidence for differential effectiveness is scarce and indications for targeted prescribing scarcer still.[57] We encourage researchers to apply the bifactor model in their analyses of within-person symptom change to tease apart the differential effects of general and specific therapeutic processes.

References

1. Angold A, Costello EJ, Erkanli, A. Comorbidity. *J Child Psychol Psychiatry.* 1999;40(1):57-87.
2. Caron C, Rutter M. Comorbidity in Child Psychopathology: Concepts, Issues and Research Strategies. *J Child Psychol Psychiatry.* 1991;32(7):1063-1080.
3. Taylor E, Rutter M. Classification: Conceptual issues and substantial findings. In: Rutter M, Taylor E. eds. *Child and adolescent psychiatry*. Oxford, England: Blackwell; 2002:3-17.
4. Achenbach TM, Ivanova MY, Rescorla LA, Turner LV, Althoff RR. Internalizing/Externalizing Problems: Review and Recommendations for Clinical and Research Applications. *J Am Acad Child Adolesc Psychiatry.* 2016;55(8):647-656.
5. Caspi A, Houts RM, Belsky DW, et al. The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders? *Clin Psychol Sci.* 2014;2(2):119-137.
6. Lahey BB, Applegate B, Hakes JK, Zald DH, Hariri AR, Rathouz PJ. Is there a general factor of prevalent psychopathology during adulthood? *J Abnorm Psychol.* 2012;121(4):971-977.
7. Caspi A, Moffitt T. All for One and One for All: Mental Disorders in One Dimension. *Am J Psychiatry*. 2018;1-14.
8. Carragher N, Teesson M, Sunderland M, et al. The structure of adolescent psychopathology: a symptom-level analysis. *Psychol Med.* 2016;46(5):981-994.
9. Deutz MH, Geeraerts SB, van Baar AL, Dekovic M, Prinzie P. The Dysregulation Profile in middle childhood and adolescence across reporters: factor structure, measurement invariance, and links with self-harm and suicidal ideation. *Eur Child Adolesc Psychiatry.* 2016;25(4):431-442.

10. Murray AL, Eisner M, Ribeaud D. The Development of the General Factor of Psychopathology 'p Factor' Through Childhood and Adolescence. *J Abnorm Child Psychol.* 2016;44(8):1573-1586.
11. Niarchou M, Moore TM, Tang SX, et al. The dimensional structure of psychopathology in 22q11.2 Deletion Syndrome. *J Psychiatr Res.* 2017;92:124-131.
12. Noordhof A, Krueger RF, Ormel J, Oldehinkel AJ, Hartman CA. Integrating autism-related symptoms into the dimensional internalizing and externalizing model of psychopathology. The TRAILS Study. *J Abnorm Child Psychol*. 2015;43(3):577-587.
13. Patalay P, Fonagy P, Deighton J, Belsky J, Vostanis P, Wolpert M. A general psychopathology factor in early adolescence. *Br J Psychiatry.* 2015;207(1):15-22.
14. St Clair MC, Neufeld S, Jones PB, et al. Characterising the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. *PLoS One.* 2017;12(4):e0175381.
15. Stochl J, Khandaker GM, Lewis G, et al. Mood, anxiety and psychotic phenomena measure a common psychopathological factor. *Psychol Med.* 2015;45(7):1483-1493.
16. Laska KM, Gurman AS, Wampold BE. Expanding the lens of evidence-based practice in psychotherapy: a common factors perspective. *Psychotherapy.* 2014;51(4):467-481.
17. Gustafsson, JE. Measurement from a hierarchical point of view. In: Braun HI, Jackson DN, Wiley DE, eds. The role of constructs in psychological and educational measurement 1st ed. London, LN: Erlbaum; 2002:73-95.
18. Fonagy P, Butler S, Cottrell D, et al. Multisystemic therapy versus management as usual in the treatment of adolescent antisocial behaviour (START): a pragmatic, randomised controlled, superiority trial. *Lancet Psychiatry.* 2018;5(2):119-133.
19. Henggeler SW, Schaeffer CM. Multisystemic Therapy((R)) : Clinical Overview, Outcomes, and Implementation Research. *Fam Process.* 2016;55(3):514-528.

20. Goodman R. The Strengths and Difficulties Questionnaire: A Research Note. *J Child Psychol Psychiatry*. 1997;38(5):581-586.
21. Goodman R. Psychometric properties of the strengths and difficulties questionnaire. *J Am Acad Child Adolesc Psychiatry.* 2001;40(11):1337-1345.
22. Muris P, Meesters C, van den Berg F. The Strengths and Difficulties Questionnaire (SDQ)--further evidence for its reliability and validity in a community sample of Dutch children and adolescents. *Eur Child Adolesc Psychiatry.* 2003;12(1):1-8.
23. Angold A, Costello EJ, Messer SC, Pickles A, Winder F, Silver D. The development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *Inter J Methods Psychiatr Res.* 1995;5:237-249.
24. Daviss WB, Birmaher B, Melhem NA, Axelson DA, Michaels SM, Brent DA. Criterion validity of the Mood and Feelings Questionnaire for depressive episodes in clinic and non-clinic subjects. *J Child Psychol Psychiatry.* 2006;47(9):927-934.
25. Wood A, Kroll L, Moore A, Harrington R. Properties of the Mood and Feelings Questionnaire in Adolescent Psychiatric Outpatients: A Research Note. *J Child Psychol Psychiatry*. 1995;36(2):327-334.
26. Lundervold AJ, Breivik K, Posserud MB, Stormark KM, Hysing M. Symptoms of depression as reported by Norwegian adolescents on the Short Mood and Feelings Questionnaire. *Front Psychol.* 2013;4:613.
27. Sharp C, Goodyer IM, Croudace TJ. The Short Mood and Feelings Questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *J Abnorm Child Psychol.* 2006;34(3):379-391.

28. Wright AG, Hopwood CJ, Skodol AE, Morey LC. Longitudinal validation of general and specific structural features of personality pathology. *J Abnorm Psychol.* 2016;125(8):1120-1134.
29. Li CH. Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res Methods.* 2016;48(3):936-949.
30. Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling.* 1999;6(1):1-55.
31. Cheung GW, Rensvold RB. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling.* 2002;9(2):233-255.
32. Murray A, Johnson W. The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence.* 2013;41(5):407-422.
33. Fabozzi F, Focardi S, Rachev S, Arshanapalli B. *The Basics of Financial Econometrics*. Hoboken, NJ: John Wiley & Sons; 2014.
34. Aitkin M, Aitkin I. Bayesian inference for factor scores. *Contemporary psychometrics: A festschrift for Roderick P. McDonald*; 2005:207-22.
35. Mislevy RJ. Randomization-based inference about latent variables from complex samples. *Psychometrika.* 1991;56(2):177-196.
36. von Davier M, Gonzalez E, & Mislevy R. What are plausible values and why are they useful. *IERI monograph series*. 2009;2:9-36.
37. Wu M. The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*. 2005;31(2):114-128.

38. Lüdtke O, Robitzsch A, Trautwein U. Integrating Covariates into Social Relations Models: A Plausible Values Approach for Handling Measurement Error in Perceiver and Target Effects. *Multivariate behavioral research*. 2018;53(1):102-124.

39. Marsman M, Maris G, Bechger T, Glas C. What can we learn from Plausible Values? *Psychometrika*. 2016;81(2):274-289.

40. Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*. 2010;92(6): 544-559.

41. Laukaityte I, Wiberg M. Using plausible values in secondary analysis in large-scale assessments. Communications in Statistics-Theory and Methods. 2017;46(22):11341-11357.

42. Skrondal A, Laake P. Regression among factor scores. *Psychometrika*. 2001;66(4):563-575.

43. Rodriguez A, Reise SP, Haviland MG. Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods.* 2016; 21(2):137–150.

44. Embretson S, Reise S. Item Response Theory for Psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2000.

45. Muthén LK, Muthén BO. Mplus User's Guide. Eighth Edition. Los Angeles, California: Muthén & Muthén; 2017.

46. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 6th ed. Essex, England: Pearson Education Limited; 2013

47. Hankin BL, Davis EP, Snyder H, Young JF, Glynn LM, Sandman CA. Temperament factors and dimensional, latent bifactor models of child psychopathology: Transdiagnostic and specific associations in two youth samples. *Psychiatry Res.* 2017;252:139-146.

48. Neumann A, Pappa I, Lahey BB, et al. Single Nucleotide Polymorphism Heritability of a General Psychopathology Factor in Children. *J Am Acad Child Adolesc Psychiatry.* 2016;55(12):1038-1045 e1034.

49. Lahey BB, Krueger RF, Rathouz PJ, Waldman ID, Zald DH. A hierarchical causal taxonomy of psychopathology across the life span. *Psychol Bull.* 2017;143(2):142-186.

50. Frick PJ, Stickle TR, Dandreaux DM, Farrell JM, Kimonis ER. Callous–Unemotional Traits in Predicting the Severity and Stability of Conduct Problems and Delinquency. *J Abnorm Child Psychol.* 2005;33(4):471-487.

51. Lahey B, Rathouz P, Keenan K, Stepp S, Loeber R, Hipwell A. Criterion validity of the general factor of psychopathology in a prospective study of girls. *J Child Psychol Psychiatry*. 2014;56(4):415-422.

52. Aseltine RH, Gore S, Gordon J. Life Stress, Anger and Anxiety, and Delinquency: An Empirical Test of General Strain Theory. *J Health Soc Behav*. 2000;41(3):256.

53. Reise S, Moore T, Maydeu-Olivares A. Target Rotations and Assessing the Impact of Model Violations on the Parameters of Unidimensional Item Response Theory Models. *Educational and Psychological Measurement*. 2011;71(4): 684–711.

54. Kim H, Eaton NR. A Hierarchical Integration of Person-Centered Comorbidity Models: Structure, Stability, and Transition Over Time. *Clinical Psychological Science*. 2017;5(4): 595–612.

55. Borsboom D. A network theory of mental disorders. *World Psychiatry*. 2017;16(1):5–13.

56. Fonagy P, Cottrell D, Phillips J, Bevington D, Glaser D, Allison E. *What works for whom?* 2nd ed. New York, NY: Guilford Press; 2015.

57. Wampold BE, Minami T, Tierney SC, Baskin TW, Bhati KS. The placebo is powerful: estimating placebo effects in medicine and psychotherapy from randomized clinical trials. *J Clin Psychol.* 2005;61(7):835-854.

Table 1. Within-level Standardized Factor Loadings for the Revised Bifactor Model With Cross-loadings.

	Factor				
Scale/Item	General	Anxiety	Antisocial	Attention	Mood
SDQ					
3. I get a lot of headaches	0.53***	0.33***			
8. I worry a lot	0.56***	0.48***			
13. I am often unhappy	0.68***	0.29***			
16. I am nervous in new situations	0.53***	0.24***	-0.32***		
24. I have many fears	0.44***	0.45***			
5. I get very angry	0.65***		0.25***		
7. I [do not] usually do as I am told	0.38***	-0.48***	0.26***		
12. I fight a lot	0.37***		0.61***		
18. I often get accused of lying or cheating	0.48***		0.33***		
22. I take things that are not mine	0.28***		0.50***		
2. I am restless	0.45***			0.68***	
10. I am constantly fidgeting	0.49***			0.66***	
15. I am easily distracted	0.54***			0.47***	
21. I [do not] think before I do things	0.45***	-0.53***		0.16***	
25. I [do not] finish the work I am doing	0.33***	-0.40***		0.16***	
MFQ					
1. I felt miserable/unhappy	0.52***				0.49***
2. I didn’t enjoy anything	0.42***				0.61***
3. I felt so tired I just sat around and did nothing	0.38***				0.49***
4. I was very restless	0.48***				0.40***
5. I felt I was no good anymore	0.63***				0.53***
M	0.48	0.40	0.38	0.43	0.50
SD	0.11	0.10	0.14	0.26	0.08
ω/ω_s	0.92	0.85	0.79	0.82	0.83

ω_H/ω_{Hs}	0.72	0.34	0.32	0.35	0.43
ECV/ ECV_S	.50	.14	.10	.12	.14

Note: ECV = Explained Common Variance; ECV_s = Explained Common Variance subscale; *M* = mean; MFQ = Mood and Feelings Questionnaire; SD = standard deviation; SDQ = Strengths and Difficulties Questionnaire; ω = Omega; ω = Omega subscale; ω_H = Omega hierarchical; ω_{Hs} = Omega hierarchical subscale.
$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

Figure 1. Total Information Functions for the General (*p*) and Specific (Antisocial, Anxiety, Attention, Mood) Psychopathology Factors

Note: Higher information (Y-axis) reflects lower standard errors hence greater reliability at different levels of the latent trait (θ; X-axis). The zero-point reflects mean factor levels.

Figure 2. Predicted and Observed Within-level Growth Curves for the General (*p*) and Specific Factors Over the Treatment and Follow-up Period

Note: Average predicted trajectories (curves) and observed means (data points with error bars) for (A) the general psychopathology and specific antisocial factors, (B) the specific anxiety factor, and (C) the specific mood and attention factors. The zero-point reflects the factor mean. Error bars indicate 95% CIs.

Supplement 1: Model Details

Our goal was to tease apart symptom-general and symptom-specific changes over a psychosocial intervention. The bifactor model is a hierarchical model designed to separate out the general and specific variance in a measure.[1] We attempted to estimate a bifactor model in addition to latent growth curves within a single-level model, but faced convergence issues. We thus split the process into two steps:

1) We first estimated the general and specific psychopathology factors at the within-level of a multilevel confirmatory bifactor analysis. This summarized how symptoms covaried over the study period for each individual.

2) We then estimated factor scores (Bayesian plausible values) of the general and specific psychopathology factors for each individual at each time-point. Factor scores were analyzed using a multilevel growth model, which included both within-person growth curves and between-person differences in within-person growth curves (i.e. random effects).

We describe the multilevel confirmatory bifactor analysis followed by the multilevel growth model in more depth below.

1) Multilevel Factor Model

We used multilevel factor analysis[2,3] to estimate within-person general and specific psychopathology factors over the study period (See Figure S1). Multilevel factor analysis is typically used to estimate separate factor structures for the within-person and between-person portions a covariance matrix. However, we used multilevel factor analysis to reduce the computational demands of estimating bifactor dimensions over time, since 'time' is treated continuously rather than discretely. In other words, a single factor can be estimated across time-points rather than repeatedly at each time-point. Data were arranged with repeated observations in long-format (e.g., vertically) and multiple items in the wide format (e.g., horizontally):

Subject	Time	Item 1	Item 2	…	Item 20
1	1	y_{11}	y_{11}		y_{11}
1	2	y_{12}	y_{12}		y_{12}
1	3	y_{13}	y_{13}		y_{13}
1	4	y_{14}	y_{14}		y_{14}
2	1	y_{21}	y_{21}		y_{21}
2	2	y_{22}	y_{22}		y_{22}
2	3	y_{23}	y_{22}		y_{23}
2	4	y_{24}	y_{24}		y_{24}
⋮					
683	4	$y_{683\ 4}$	$y_{683\ 4}$		$y_{683\ 4}$

Each item was specified at the within-level (level 1). We did not allow for variances at the between-level (level 2), but corrected the standard errors for the nesting of observations within subjects using a subject ID cluster variable. The model can be expressed as follows:

$$Y_{ijt} = v_{W_{ijt}} + \Lambda_{\mathrm{W}}\eta_{\mathrm{W}_{ijt}} + \varepsilon_{\mathrm{W}_{ijt}}$$

where Y is a matrix reflecting the observed responses on each item, $j = 1,\ldots,J$, at each time-point, $t = 1,\ldots,T$ across individuals, $i = 1,\ldots,N$, $v_{W_{ijt}}$ is a vector of within-level item thresholds; Λ_{W} is a within-level factor loading matrix, $\eta_{\mathrm{W}_{ijt}}$ is a vector of factors which vary randomly across time-points and items within subjects, and e_{ijt} is the within-person error. The $\Lambda_{\mathrm{W}}\eta_{\mathrm{W}_{ijt}}$ term can be expressed more fully as:

$$\Lambda_{\mathrm{W}}\eta_{\mathrm{W}_{ijt}} = \lambda_{Wgeneral_j}\theta_{Wgeneral_{it}} + \lambda_{Wspecific1_j}\theta_{Wspecific1_{it}} + \lambda_{Wspecific2_j}\theta_{Wspecific2_{it}} + \lambda_{Wspecific3_j}\theta_{Wspecific3_{it}} + \lambda_{Wspecific4_j}\theta_{Wspecific4_{it}}$$

where λ_{W_j} are within-level factor loadings for each item and $\theta_{W_{it}}$ are within-level factor vectors which vary across subjects and time-points for the general factor, $general$, and specific factors, $specific1, \ldots, specificK$, where $K = 4$ in the current model.

Our notation implies that this was a three-level model, with repeated observations at the lowest level ('time') nested in each item ('item'), nested within individuals ('subject'). However, when implementing the model in Mplus, we included each item as a different within-level variable (see the data structure table above), making it a multi-indicator two-level factor model. Nonetheless, the models are equivalent.

2) Multilevel Growth Model

We estimated Bayesian plausible values (i.e. a distribution of factor scores) for the general and specific within-level factors described above. We thus had several estimates of each subject's score on each factor at each time-point (e.g., $\hat{\theta}_{it}$), which were averaged over using multiple imputation. For simplicity, we refer to a single set of factor scores. Data were formatted with repeated observations for each factor in long format (e.g., vertically) and each factor in wide format (e.g., horizontally):

Subject	Time	θ_p	$\theta_{antisocial}$	$\theta_{anxiety}$	$\theta_{attention}$	θ_{mood}
1	0	y_{10}	y_{10}	y_{10}	y_{10}	y_{10}
1	1	y_{11}	y_{11}	y_{11}	y_{11}	y_{11}
1	2	y_{12}	y_{12}	y_{12}	y_{12}	y_{12}
1	3	y_{13}	y_{13}	y_{13}	y_{13}	y_{13}
2	0	y_{20}	y_{20}	y_{20}	y_{20}	y_{20}
2	1	y_{21}	y_{21}	y_{21}	y_{21}	y_{21}
2	2	y_{22}	y_{22}	y_{22}	y_{22}	y_{22}
2	3	y_{23}	y_{23}	y_{23}	y_{23}	y_{23}
⋮						
683	3	$y_{683\ 3}$	$y_{683\ 3}$	$y_{683\ 3}$	$y_{683\ 3}$	$y_{683\ 3}$

We estimated a two-level parallel process growth model using factor scores as outcome variables (See Figure S2). The simultaneous analysis of growth in each factor, $f = 1,\ldots,F$, is denoted with a superscript (items in the multilevel factor model described above were also analyzed simultaneously, but denoted with a subscript). The within-level or level 1 portion of the model can be written as:

$$y_{it}^{(f)} = \beta_{0_i}^{(f)} + \beta_{1_i}^{(f)} Time_{it} + \beta_{2_i}^{(f)} Time^2{}_{it} + \varepsilon_{it}^{(f)}$$

where $y_{it}^{(f)}$ reflects factor scores for each individual, $i = 1,\ldots,N$ at each time-point, $t = 0,\ldots,T$ for a given factor, $\beta_{0_i}^{(f)}$ reflects the intercept or baseline factor scores for each individual when $t = 0$ (for each factor), $\beta_{1_i}^{(f)}$ and $\beta_{2_i}^{(f)}$ reflect the linear and quadratic slopes of time on each factor, respectively, which vary randomly across individuals, $Time_{it}$ and $Time^2{}_{it}$ reflect the observed values of time (0, 1, 2, 3) and time-squared (0, 1, 4, 9) for each individual at each time-point; and $\varepsilon_{it}^{(f)}$ reflects the individual- and time-specific residuals.

The between-level or level 2 part of the model can be expressed as

$$\beta_{0_i}^{(f)} = \gamma_{00}^{(f)} + \gamma_{01}^{(f)} c.Age_i + U_{0i}^{(f)}$$
$$\beta_{1_i}^{(f)} = \gamma_{10}^{(f)} + \gamma_{11}^{(f)} c.Age_i + U_{1i}^{(f)}$$
$$\beta_{2_i}^{(f)} = \gamma_{20}^{(f)} + \gamma_{21}^{(f)} c.Age_i + U_{2i}^{(f)}$$

where $\gamma_{00}^{(f)}$, $\gamma_{10}^{(f)}$, and $\gamma_{20}^{(f)}$ are the overall mean intercept, mean linear slope of time, and mean quadratic slope of time, respectively, across individuals for each factor; $\gamma_{01}^{(f)}$,$\gamma_{11}^{(f)}$, and $\gamma_{21}^{(f)}$ are the effect of between-person differences in baseline age (centred) on the intercept, linear time slope, and quadratic time slope for each factor, respectively; $c.Age_i$ reflects each person's baseline age centred using the sample mean age at baseline; and $U_{0i}^{(f)}$, $U_{1i}^{(f)}$, and $U_{2i}^{(f)}$ reflect person-specific deviations from the overall intercept, linear slope of time, and quadratic slope of time, respectively, for each factor.

The covariance structure for the random effects across factors was unrestricted. That is, we freely estimated the covariances between the random intercepts, linear slopes, and quadratic slopes for each factor, as well as between factors, forming a 15 x 15 unrestricted covariance matrix:

$$V\begin{bmatrix} U_{0i}^{(1)} \\ U_{1i}^{(1)} \\ U_{2i}^{(1)} \\ U_{0i}^{(2)} \\ U_{1i}^{(2)} \\ U_{2i}^{(2)} \\ \vdots \\ U_{2i}^{(5)} \end{bmatrix} = \begin{bmatrix} \tau_{00}^{(1)} & & & & & & & \\ \tau_{10}^{(1)} & \tau_{11}^{(1)} & & & & & & \\ \tau_{20}^{(1)} & \tau_{21}^{(1)} & \tau_{22}^{(1)} & & & & & \\ \tau_{00}^{(2,1)} & \tau_{01}^{(2,1)} & \tau_{02}^{(2,1)} & \tau_{00}^{(2)} & & & & \\ \tau_{10}^{(2,1)} & \tau_{11}^{(2,1)} & \tau_{12}^{(2,1)} & \tau_{10}^{(2)} & \tau_{11}^{(2)} & & & \\ \tau_{20}^{(2,1)} & \tau_{21}^{(2,1)} & \tau_{22}^{(2,1)} & \tau_{20}^{(2)} & \tau_{21}^{(2)} & \tau_{22}^{(2)} & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ \tau_{20}^{(5,1)} & \tau_{21}^{(5,1)} & \tau_{22}^{(5,1)} & \tau_{20}^{(5,2)} & \tau_{21}^{(5,2)} & \tau_{22}^{(5,2)} & \ldots & \tau_{22}^{(5)} \end{bmatrix}$$

References

1. Reise SP. The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research.* 2012;47(5):667–696.
2. Muthén BO. Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement.* 1991;28(4):338-354.
3. Muthén BO. Multilevel covariance structure analysis. *Sociological methods & research.* 1994;22(3):376-398.

Supplement 2: Longitudinal Measurement Invariance Testing

We used multilevel factor analysis with 'time' at the within-level and 'subject' at the between-level to estimate general and specific psychopathology factors at the within-level over time (see Supplement 1). A disadvantage of this modeling approach is that it was not possible to test for measurement invariance in the conventional sense, i.e. by holding factor loadings and item intercepts/thresholds constant at each time-point. This is because 'time' is an inherent feature of model parameters, e.g., a within-level factor loading reflects the way in which an item is predicted to covary with other items across time. In contrast, the conventional measurement invariance test relies on a single-level model, where factors are estimated at each time-point, and hence model parameters can be freely estimated or held constant at each time-point. In the multilevel approach, factor loadings and item intercepts/thresholds are assumed to be invariant. For example, an item intercept is the mean of that item over the within-level (e.g., time) when a given factor equals zero.

The reviewers and authors agreed that some type of invariance testing should be undertaken to support the assumption that change was mainly attributable to the factors and not changes in measurement properties. This is despite the fact that full or partial measurement invariance shown with the conventional approach would demonstrate properties of the parameters that are not immediately transferable to the multi-level approach. A factor loading in one model is not the same as a factor loading in the other. Moreover, full or partial invariance shown using the conventional approach cannot be carried over to the multilevel model, since there are simply no parameters to hold constant. That said, the results of both single-level and multi-level growth models should ultimately converge, and so invariance observed using one approach should roughly translate to the other.

We encountered convergence issues when estimating a single-level model with wide-formatted data. We believe this was mainly due to model complexity (e.g., simultaneously estimating four bifactor models in addition to growth factors is computationally taxing). We thus estimated the general and specific psychopathology factors for two adjacent time-points within the same single-level model, which converged successfully. However, when we attempted to assess metric invariance (e.g., equal factor loadings between the adjacent time-points), chi-square difference values between models were negative, which is possible but improper and non-meaningful.[1]

As an alternative, we tested the invariance of individual factor loadings between two adjacent time-points using Wald chi-square tests via the MODEL CONSTRAINT command in Mplus. We found that all factor loadings showed metric invariance except for those associated with the mood factor between time 2 (post-treatment) and time 3 (6-months follow-up), Wald $\chi^2(4) = 11.54$, $p = .021$ (the Wald test includes all mood items for brevity but each item was initially tested individually).

We then tested for scalar invariance by comparing individual item thresholds between two adjacent time-points using Wald chi-square tests, while simultaneously testing for differences among all factor loadings (the latter was intended to mimic equality constraints on all factor loadings, which is a prerequisite when testing scalar invariance). Each of the 20 items had two thresholds (threshold A and B) which were compared at three adjacent time-points (time 1 vs. time 2, time 2 vs. time 3, time 3 vs. time 4), resulting in 120 tests. To minimize family wise error rates, we corrected the alpha level for the number of tests conducted on a single threshold between two adjacent time-points using the Bonferroni method (e.g., α/k, where α

is the type I error rate and *k* is the number of tests). Therefore, $\alpha = .003$ ($\alpha/k = .05/20$) when testing the equivalence of one of the two thresholds for each of the 20 items between two adjacent time-points.

Threshold A was invariant for 80% of items between time 1 and 2, while threshold B was invariant for 60% of items. Between time 2 and 3, threshold A was invariant for 90% of items, while threshold B was invariant for 95% of items. Finally, 100% of items showed invariance in threshold A and B between time 3 and 4. Non-invariance of item thresholds was thus mainly apparent between time 1 (baseline) and 2 (post-treatment), which may be because pre-treatment distributions can deviate from post-treatment distributions.[2,3] Three of the nine items (33%) that showed non-invariance in threshold A between time 1 and 2 also showed non-invariance in threshold B (e.g., SDQ items 5 and 12, and MFQ item 5). Therefore, the majority of non-invariance appeared sporadic rather than systematic.

In all, our conventional measurement invariance analysis demonstrates partial longitudinal measurement invariance, but caution is warranted when extending these findings to the multilevel model.

References

1. Satorra A, Bentler P. M. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika.* 2001;66(4):507-514.
2. Hedeker D, Gibbons R. D. *Longitudinal Data Analysis*. 2006; John Wiley & Sons, Hoboken, NJ.
3. Vickers A. J. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC medical research methodology*. 2005;5(1)5-35.

Supplement 3: Sensitivity Analysis – Growth Model Without Cross-loadings

We re-ran the multilevel growth model described in the paper using Bayesian plausible values from a bifactor model that did not include cross-loadings (see paper for model fit and Table S5 for factor loadings). Our goal was to determine the influence of cross-loadings on the direction and significance of the growth curves, particularly for the specific anxiety and antisocial factors. The decline in antisocial scores may have been driven by an increase in the negatively weighted anxiety item which cross-loaded onto the antisocial factor. Similarly, anxiety scores may have increased because of a decrease in the negatively weighted antisocial item or attention items which cross-loaded.

In the multilevel growth model without cross-loadings, the anxiety factor continued to show a significant linear increase over the study period (β = .34, $p < .001$, 95% CI [.18, .51]; see Figure S3b). The increase was stronger in magnitude than the model that included cross-loadings, most likely because of SDQ item 16's boost in loading strength from no longer cross-loading on the antisocial factor. Overall, it does not appear that the antisocial and attention items that cross-loaded on the anxiety factor underpinned its increase over time.

In contrast, the antisocial factor still declined over the study period (β = -.05, $p = .614$, 95% CI [-.22, .13]) but at a weaker magnitude which was no longer significant (see Figure S3a). Hence, it appears that the negatively weighted SDQ item 16 ('I am [not] nervous in new situations') contributed much to the decline in antisocial scores. However, to say that antisocial scores declined because of an increase in anxiety may not be entirely accurate, because SDQ item 16 loaded more strongly onto, and hence better represents, the antisocial factor than the anxiety factor. We would argue that in the context of the antisocial factor, SDQ item 16 reflects fearlessness more than separation anxiety (the original item meaning). Furthermore, forcing SDQ item 16 to load exclusively onto the anxiety factor despite its affinity to the antisocial factor may have supressed the latter's growth curve in the parallel process growth model.

As for the other factors, the p factor continued to decline over time (β = -.47, $p < .001$, 95% CI [-.60, -.34]), which, like the anxiety factor, was stronger in magnitude than the model featuring cross-loadings (see Figure S3a). Removing the cross-loadings appears to have strengthened changes in the general variance**,** perhaps because the general factor may absorb the variance associated with unmodelled cross-loadings.[1] Moreover, the quadratic slope for the p factor was now significant, albeit just (β = .04, $p = .045$, 95% CI [.01, .08]). The mood (β = -.04, $p = .638$, 95% CI [-.21, .13]) and attention (β = .02, $p = .779$, 95% CI [-.12, .16]) factors both decreased slightly in their baseline values compared to the model with cross-loadings, but continued to show little change over time (see Figure S3c).

References

1. Murray A, Johnson W. The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. Intelligence. 2013;41(5):407-422.

Table S1. Standardized Factor Loadings for the Mood and Feelings Questionnaire (Exploratory Within-level Factor Analysis)

	Factor	
Scale/Item	Self-Attitudes	Mood
1. I felt miserable or unhappy.	0.35	**0.36**
2. I didn't enjoy anything at all.	0.36	**0.34**
3. I felt so tired I just sat around and did nothing.	0.02	**0.61**
4. I was very restless.	-0.01	**0.68**
5. I felt I was no good anymore.	0.69	**0.34**
6. I cried a lot.	0.65	0.14
7. I found it hard to think properly or concentrate.	0.38	0.26
8. I hated myself.	0.85	0.02
9. I was a bad person.	0.72	0.00
10. I felt lonely.	0.78	0.03
11. I thought nobody really loved me.	0.84	-0.02
12. I thought I could never be as good as other kids.	0.83	-0.08
13. I did everything wrong.	0.81	-0.05

Note: Top five items loading ≥ .32 on the mood factor are in bold and were used in the primary model.

Table S2. Correlation Matrix of Bayesian Plausible Values for the General (*p*) and Specific (Anxiety, Mood, Antisocial, Attention) Psychopathology Factors

	p	Anxiety	Mood	Antisocial	Attention
p	—				
Anxiety	-0.042	—			
Mood	0.002	0.048	—		
Antisocial	0.06	-0.079	-0.004	—	
Attention	0.003	-0.016	-0.025	0.034	—

Note: The average number of observations over 100 imputations was 2,732 for 683 cases. Correlations between factors were set at zero in the original model.

Table S3. Within-level Polychoric Correlation Matrix. Items are Arranged by Specific Factor (eg, 1-5 = Anxiety, 6-10 = Mood, 11-15 = Antisocial, and 16-20 = Attention)

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
1. SDQ 3	—																			
2. SDQ 8	0.43	—																		
3. SDQ 13	0.45	0.56	—																	
4. SDQ 16	0.29	0.44	0.39	—																
5. SDQ 24	0.40	0.51	0.42	0.40	—															
6. MFQ 1	0.39	0.43	0.55	0.29	0.33	—														
7. MFQ 2	0.26	0.25	0.40	0.21	0.22	0.54	—													
8. MFQ 3	0.30	0.26	0.29	0.21	0.19	0.38	0.45	—												
9. MFQ 4	0.30	0.25	0.32	0.18	0.18	0.42	0.39	0.50	—											
10. MFQ 5	0.36	0.44	0.57	0.29	0.32	0.63	0.61	0.45	0.46	—										
11. SDQ 5	0.29	0.24	0.34	0.23	0.17	0.35	0.25	0.24	0.30	0.37	—									
12. SDQ 7	0.01	0.01	0.18	0.00	-0.01	0.11	0.18	0.08	0.17	0.20	0.36	—								
13. SDQ 12	0.19	0.04	0.28	-0.01	0.05	0.15	0.17	0.11	0.22	0.20	0.43	0.29	—							
14. SDQ 18	0.23	0.19	0.29	0.16	0.13	0.24	0.23	0.16	0.27	0.28	0.35	0.25	0.35	—						
15. SDQ 22	0.08	0.03	0.24	0.03	0.11	0.12	0.14	0.03	0.16	0.21	0.21	0.24	0.41	0.39	—					
16. SDQ 2	0.25	0.20	0.23	0.24	0.14	0.20	0.13	0.19	0.40	0.22	0.41	0.15	0.27	0.28	0.14	—				
17. SDQ 10	0.27	0.22	0.25	0.30	0.17	0.18	0.15	0.17	0.36	0.24	0.39	0.18	0.28	0.31	0.17	0.68	—			
18. SDQ 15	0.24	0.26	0.24	0.38	0.17	0.17	0.14	0.19	0.29	0.24	0.46	0.24	0.24	0.31	0.15	0.55	0.57	—		
19. SDQ 21	0.07	0.03	0.14	0.07	0.02	0.11	0.14	0.10	0.18	0.22	0.38	0.43	0.26	0.27	0.21	0.29	0.31	0.37	—	
20. SDQ 25	0.02	0.04	0.11	0.18	0.00	0.11	0.11	0.06	0.14	0.17	0.20	0.37	0.09	0.13	0.10	0.22	0.23	0.34	0.41	—

Table S4. Within-Level Standardized Factor Loadings for the Common Factor Model

Scale/Item	Factor General
SDQ	
3. I get a lot of headaches	0.52***
8. I worry a lot	0.55***
13. I am often unhappy	0.65***
16. I am nervous in new situations	0.46***
24. I have many fears	0.45***
5. I get very angry	0.61***
7. I [do not] usually do as I am told	0.34***
12. I fight a lot	0.40***
18. I often get accused of lying or cheating	0.48***
22. I take things that are not mine	0.31***
2. I am restless	0.61***
10. I am constantly fidgeting	0.64***
15. I am easily distracted	0.62***
21. I [do not] think before I do things	0.41***
25. I [do not] finish the work I am doing	0.31***
MFQ	
1. I felt miserable/unhappy	0.63***
2. I didn't enjoy anything	0.57***
3. I felt so tired I just sat around and did nothing	0.49***
4. I was very restless	0.58***
5. I felt I was no good anymore	0.73***
M	0.52
SD	0.12

Note: *M* = mean; MFQ = Mood and Feelings Questionnaire; SDQ = Strengths and Difficulties Questionnaire.

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table S5. Within-Level Standardized Factor Loadings for the Correlated Factors Model and Factor Correlations

	Factor			
Scale/Item	Anxiety	Antisocial	Attention	Mood
SDQ				
3. I get a lot of headaches	0.63***			
8. I worry a lot	0.70***			
13. I am often unhappy	0.80***			
16. I am nervous in new situations	0.57***			
24. I have many fears	0.57***			
5. I get very angry		0.78***		
7. I [do not] usually do as I am told		0.46***		
12. I fight a lot		0.54***		
18. I often get accused of lying or cheating		0.60***		
22. I take things that are not mine		0.42***		
2. I am restless			0.74***	
10. I am constantly fidgeting			0.78***	
15. I am easily distracted			0.76***	
21. I [do not] think before I do things			0.54***	
25. I [do not] finish the work I am doing			0.42***	
MFQ				
1. I felt miserable/unhappy				0.74***
2. I didn't enjoy anything				0.67***
3. I felt so tired I just sat around and did nothing				0.58***
4. I was very restless				0.64***
5. I felt I was no good anymore				0.86***
M	0.65	0.70	0.56	0.65
SD	0.10	0.15	0.14	0.16
	1.	2.	3.	4.

1. Anxiety	—			
2. Antisocial	0.43***	—		
3. Attention	0.43***	0.72***	—	
4. Mood	0.69***	0.52***	0.39***	—

Note: *M* = mean; MFQ = Mood and Feelings Questionnaire; *SD* = standard deviation; SDQ = Strengths and Difficulties Questionnaire.
*$p < .05$; **$p < .01$; ***$p < .001$

Table S6. Within-Level Standardized Factor Loadings for a Confirmatory Bifactor Model Without Cross-loadings

	Factor				
Scale/Item	General	Anxiety	Antisocial	Attention	Mood
SDQ					
3. I get a lot of headaches	0.49***	0.34***			
8. I worry a lot	0.46***	0.63***			
13. I am often unhappy	0.62***	0.40***			
16. I am nervous in new situations	0.42***	0.38***			
24. I have many fears	0.34***	0.59***			
5. I get very angry	0.67***		0.22***		
7. I [do not] usually do as I am told	0.35***		0.29***		
12. I fight a lot	0.37***		0.57***		
18. I often get accused of lying or cheating	0.48***		0.35***		
22. I take things that are not mine	0.27***		0.55***		
2. I am restless	0.47***			0.64***	
10. I am constantly fidgeting	0.51***			0.63***	
15. I am easily distracted	0.55***			0.48***	
21. I [do not] think before I do things	0.39***			0.27***	
25. I [do not] finish the work I am doing	0.28***			0.28***	
MFQ					
1. I felt miserable/unhappy	0.54***				0.47***
2. I didn’t enjoy anything	0.45***				0.60***
3. I felt so tired I just sat around and did nothing	0.41***				0.47***
4. I was very restless	0.52***				0.35***
5. I felt I was no good anymore	0.66***				0.50***
M	0.46	0.47	0.40	0.46	0.48
SD	0.11	0.13	0.16	0.18	0.09
ω/ω_s	0.91	0.80	0.73	0.78	0.83

ω_H/ω_{Hs}	0.73	0.40	0.34	0.41	0.39
ECV/ECV$_s$	0.51	0.13	0.10	0.13	0.13

Note: ECV = Explained Common Variance; ECV$_s$ = Explained Common Variance subscale; *M* = mean; MFQ = Mood and Feelings Questionnaire; SD = standard deviation; SDQ = Strengths and Difficulties Questionnaire; ω = omega; ω$_s$ = omega subscale; ω_H = omega hierarchical; ω_{Hs} = omega hierarchical subscale.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table S7. Within-Level Standardized Factor Loadings for an Exploratory Bi-factor Model (Bi-Geomin Orthogonal Rotation)

	Factor				
Scale/Item	General	Anxiety	Antisocial	Attention	Mood
SDQ					
3. I get a lot of headaches	0.55	0.26	-0.07	0.01	0.09
8. I worry a lot	0.65	0.44	-0.30	-0.14	-0.03
13. I am often unhappy	0.74	0.17	0.00	-0.23	0.08
16. I am nervous in new situations	0.54	0.33	**-0.37**	0.06	-0.08
24. I have many fears	0.54	0.39	-0.24	-0.13	-0.07
5. I get very angry	0.58	-0.16	0.25	0.16	0.04
7. I [do not] usually do as I am told	0.30	**-0.50**	0.32	-0.04	0.02
12. I fight a lot	0.38	0.02	0.60	0.11	-0.03
18. I often get accused of lying or cheating	0.46	-0.03	0.33	0.09	0.01
22. I take things that are not mine	0.32	-0.01	0.48	-0.04	-0.07
2. I am restless	0.48	-0.04	0.05	0.66	0.05
10. I am constantly fidgeting	0.52	-0.06	0.04	0.62	-0.02
15. I am easily distracted	0.56	-0.25	-0.04	0.44	-0.10
21. I [do not] think before I do things	0.35	**-0.52**	0.23	0.14	-0.01
25. I [do not] finish the work I am doing	0.28	**-0.58**	-0.03	0.09	-0.03
MFQ					
1. I felt miserable/unhappy	0.60	0.08	-0.05	-0.20	0.39
2. I didn’t enjoy anything	0.46	-0.04	0.03	-0.18	0.54
3. I felt so tired I just sat around and did nothing	0.37	0.05	-0.07	0.06	0.54
4. I was very restless	0.45	0.02	0.07	0.23	0.51
5. I felt I was no good anymore	0.66	-0.04	0.00	-0.21	0.46

Note: Items in bold reflect cross-loadings meeting the threshold of .32. Model fit: CFI = .95, TLI = .91, RMSEA = .06, SRMR = .04. MFQ = Mood and Feelings Questionnaire; SDQ = Strengths and Difficulties Questionnaire.

Table S8. Correlations Between Random Intercepts, Random Linear Slopes, and Random Quadratic Slopes for the General (p) and Specific Psychopathology Factors

	1.	2.	3.	4.	5.	6	7.	8.	9.	10.	11.	12.	13.	14.	15.
1. $U_{0i}^{(p)}$	0.38***														
2. $U_{1i}^{(p)}$	-0.11	0.26													
3. $U_{2i}^{(p)}$	0.02	-0.08	0.03												
4. $U_{0i}^{(anxiety)}$	-0.02	0.02	0.00	0.22***											
5. $U_{1i}^{(anxiety)}$	0.07	-0.07	0.02	-0.14	0.27										
6. $U_{2i}^{(anxiety)}$	-0.01	0.02	-0.01	0.03	-0.08	0.03									
7. $U_{0i}^{(mood)}$	-0.01	0.07	-0.02	0.06	-0.05	0.01	0.16*								
8. $U_{1i}^{(mood)}$	0.09	-0.14	0.04	-0.04	0.09	-0.02	-0.14	0.27							
9. $U_{2i}^{(mood)}$	-0.02	0.04	-0.01	0.01	-0.02	0.01	0.03	-0.08	0.03						
10. $U_{0i}^{(anti)}$	-0.01	0.05	-0.01	-0.03	-0.01	0.00	-0.01	0.00	0.00	0.16*					
11. $U_{1i}^{(anti)}$	0.04	-0.07	0.02	-0.01	0.02	-0.01	0.03	-0.01	0.00	-0.14	0.31				
12. $U_{2i}^{(anti)}$	-0.01	0.02	-0.01	0.01	-0.01	0.00	-0.01	0.00	0.00	0.03	-0.09	0.03			
13. $U_{0i}^{(atten)}$	0.02	0.06	-0.02	-0.04	0.02	0.00	-0.03	0.02	-0.01	0.02	-0.03	0.01	0.24***		
14. $U_{1i}^{(atten)}$	0.00	-0.11	0.04	-0.01	0.05	-0.02	0.00	-0.01	0.00	0.00	0.05	-0.02	-0.17	0.33	
15. $U_{2i}^{(atten)}$	0.00	0.03	-0.01	0.00	-0.02	0.01	0.00	0.00	0.00	0.00	-0.01	0.01	0.04	-0.09	0.03

Note: Variances are on the diagonal. anti = specific antisocial factor; atten = specific attention factor; p = general psychopathology; U_{0i} = random intercept; U_{1i} = random linear slope; U_{2i} = random quadratic slope.

$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

Table S9. Regression Coefficients of the Random Effects for Each Factor on Baseline Age

Parameter	*B*	*p*	95% LL	95% UP
Random Intercept				
p	-0.03	0.24	-0.09	0.02
Anxiety	0.02	0.57	-0.05	0.08
Mood	-0.02	0.58	-0.08	0.04
Antisocial	-0.02	0.62	-0.09	0.05
Attention	-0.02	0.46	-0.09	0.04
Random Slope (Linear)				
p	0.06	0.16	-0.02	0.15
Anxiety	0.00	0.99	-0.11	0.11
Mood	0.03	0.62	-0.08	0.13
Antisocial	-0.02	0.73	-0.14	0.10
Attention	-0.02	0.74	-0.10	0.07
Random Slope (Quadratic)				
p	-0.02	0.11	-0.05	0.00
Anxiety	0.00	0.99	-0.03	0.03
Mood	0.00	0.82	-0.04	0.03
Antisocial	0.01	0.71	-0.03	0.04
Attention	0.01	0.64	-0.02	0.03

Note: *B* = partially standardized beta; LL = lower limit; UP = upper limit

Figure S1. Schematic of the Item-Level Multilevel Confirmatory Bi-factor Analysis With Cross-loadings

Note: Each box reflects an observed item from the Strengths and Difficulties Questionnaire (SDQ) or Mood and Feelings Questionnaire (MFQ). Each circle reflects a latent variable which was estimated at the within-level only. p = general psychopathology; Anx = anxiety; Anti = antisocial; Atten = attention.

Figure S2. Schematic of the Multilevel Growth Curve Model Using Bayesian Plausible Values for the Within-level Bifactor Dimensions

Note: General (p) and specific psychopathology factor scores were regressed onto linear and quadratic time variables. Random effects are illustrated by the black circles at the end of the path (random intercepts) and at the middle of the path labelled with an S (random slopes). At the between level, the random intercept (i), random linear slope (s), and random quadratic slope (s^2) for each factor were correlated, and also regressed on a centered age variable. p = general psychopathology; Anx = anxiety; Anti = antisocial; Atten = attention; c.Age = age centred.

Figure S3. Predicted and Observed Within-level Growth Curves for the p Factor and Specific Anxiety, Mood, Antisocial, and Attention Factor BPVs Estimated From a Model Without Cross-loadings.

Note: Average predicted trajectories (curves) and observed means (data points with error bars) for (A)the general psychopathology and specific antisocial factors, (B) the specific anxiety factor, and (C) the specific mood and attention factors. The zero-point reflects the factor mean. Error bars indicate 95% confidence intervals.